Feature-space transform tying in unified acoustic-articulatory modelling of articulatory control of HMM-based speech synthesis

Feature-Space Transform Tying in Unified Acoustic-Articulatory Modelling for Articulatory Control of HMM-based Speech Synthesis

Zhen-Hua Ling[1], *Korin Richmond*[2], *Junichi Yamagishi*[2]

[1]iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China
[2]CSTR, University of Edinburgh, United Kingdom

zhling@ustc.edu, korin@cstr.ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

In previous work, we have proposed a method to control the characteristics of synthetic speech flexibly by integrating articulatory features into hidden Markov model (HMM) based parametric speech synthesis. A unified acoustic-articulatory model was trained and a piecewise linear transform was adopted to describe the dependency between these two feature streams. The transform matrices were trained for each HMM state and were tied based on each state's context. In this paper, an improved acoustic-articulatory modelling method is proposed. A Gaussian mixture model (GMM) is introduced to model the articulatory space and the cross-stream transform matrices are trained for each Gaussian mixture instead of context-dependently. This means the dependency relationship can vary with the change of articulatory features flexibly. Our results show this method improves the effectiveness of control over vowel quality by modifing articulatory trajectories without degrading naturalness.

Index Terms: speech synthesis, articulatory features, hidden Markov model, Gaussian mixture model

1. Introduction

The hidden Markov model (HMM)-based parametric speech synthesis method has made significant progress in recent years [1, 2]. This method is able to synthesize highly intelligible and smooth speech sounds [3, 4]. In our previous work, we have proposed a method to improve the flexibility of HMM-based speech synthesis by integrating articulatory features [5,6]. Here, we use "articulatory features" to refer to the continuous movements of a group of speech articulators, such as the tongue, jaw, lips and velum, recorded by human articulography techniques. In this method, a unified acoustic-articulatory model with cross-stream dependency is trained. During synthesis, the characteristics of synthetic speech can be controlled flexibly by modifying the generated articulatory features according to arbitrary phonetic rules. Experimental results have shown the effectiveness of this method in controlling the overall character of synthesized speech and the quality of specific vowels [6].

A piecewise linear transform was used to describe the dependency of acoustic feature production on the movement of articulatory features in our previous work [5, 6]. Like other model parameters in the unified acoustic-articulatory HMM, the transform matrices were trained for each HMM state and were tied based on context using a decision tree. Therefore, the cross-stream dependency was entirely determined by the context information of input text. This could become problematic when the articulatory features were modified using phonetic rules during synthesis because the transform matrix was expected to adapt to the new articulatory configuration. In this paper, a feature-space transform tying method is proposed to solve this problem. A Gaussian mixture model (GMM) is adopted to model the articulatory space and the cross-stream transform matrices are estimated for each Gaussian component instead of for each HMM state (thus depending on its context information).

This paper is organized as follows. Section 2 gives a brief overview of our baseline acoustic-articulatory modelling method. Section 3 describes our proposed method in detail. Section 4 introduces the results of our experiments and Section 5 presents the conclusions we draw from this work.

2. Baseline

In our baseline method, the general framework of HMM-based speech synthesis was followed to integrate articulatory features into the conventional modelling of acoustic features [6]. Let $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_T^\top]^\top$ denote the parallel acoustic and articulatory feature sequence of the same length T. For each frame, the feature vector $\boldsymbol{x}_t \in \mathcal{R}^{3D_X}$ and $\boldsymbol{y}_t \in \mathcal{R}^{3D_Y}$ consist of static parameters and their velocity and acceleration components, where D_X and D_Y are the dimensions of static acoustic features and static articulatory features respectively. In model training, an HMM λ is estimated by maximizing the likelihood function of the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}|\lambda)$. A piecewise (state-wise) linear transform is added to the model parameters to represent the dependency between the generation of acoustic features and the articulatory movements. The joint distribution can be written as

$$P(\boldsymbol{X}, \boldsymbol{Y}|\lambda) = \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t, \boldsymbol{y}_t), \quad (1)$$

$$b_j(\boldsymbol{x}_t, \boldsymbol{y}_t) = b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) b_j(\boldsymbol{y}_t), \quad (2)$$

$$b_j(\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{Y_j}, \boldsymbol{\Sigma}_{Y_j}), \quad (3)$$

$$b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_j \boldsymbol{y}_t + \boldsymbol{\mu}_{X_j}, \boldsymbol{\Sigma}_{X_j}). \quad (4)$$

where $\boldsymbol{q} = \{q_1, q_2, ..., q_N\}$ is the state sequence shared by the two feature streams; π_j and a_{ij} represent initial state probability and state transit probability; $b_j(\cdot)$ is the state observation probability density function (PDF) for state j; $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$; $\boldsymbol{A}_j \in \mathcal{R}^{3D_X \times 3D_Y}$ is the linear transform matrix for state j. This matrix is context-dependent, hence a globally piecewise linear transform can be achieved. The model parameters can be estimated using the EM algorithm [6].

During synthesis, the acoustic and articulatory features are simultaneously generated from the trained models using maximum-likelihood parameter generation (MLPG) algorithm that considers explicit constraints of the dynamic features. In order to control the characteristics of synthetic speech flexibly, the generated articulatory features can be modified based

on phonetic knowledge to reproduce acoustic parameters that reflect those changes appropriately [6].

3. Proposed Method

3.1. Model Structure

As discussed above, the transform matrix $\boldsymbol{A}_j$ in Eq.(4) is tied across states according to context information. This may lead to the incorrect representation of feature dependency when the generated articulatory features are modified during synthesis. In this paper, we improve the model structure so that the transform matrix can be determined by the articulatory features instead of the context information. Here, a GMM model $\lambda^{(G)}$ of M mixtures is trained in advance using only the articulatory stream of training data to represent the articulatory space. Then, the transform matrices are trained for each mixture component of $\lambda^{(G)}$. Mathematically, we rewrite Eq.(4) as

$$b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) = \sum_{k=1}^{M} P(\boldsymbol{x}_t, m_t = k|\boldsymbol{y}_t, q_t = j, \lambda, \lambda^{(G)}), \quad (5)$$

$$= \sum_{k=1}^{M} \zeta_k(t) P(\boldsymbol{x}_t|\boldsymbol{y}_t, q_t = j, m_t = k, \lambda, \lambda^{(G)}), \quad (6)$$

where m_t denotes the mixture index of $\lambda^{(G)}$ for articulatory feature vector at frame t; the HMM state sequence $\boldsymbol{q}$ and the GMM mixture sequence $\boldsymbol{m} = \{m_1, m_2, ..., m_N\}$ are assumed to be independent, i.e.

$$P(m_t = k|\boldsymbol{y}_t, q_t = j, \lambda, \lambda^{(G)}) = P(m_t = k|\boldsymbol{y}_t, \lambda^{(G)}) = \zeta_k(t). \quad (7)$$

For each Gaussian mixture, the dependency between the acoustic and articulatory features is represented as

$$P(\boldsymbol{x}_t|\boldsymbol{y}_t, q_t = j, m_t = k, \lambda, \lambda^{(G)}) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_k \boldsymbol{\xi}_t + \boldsymbol{\mu}_{X_j}, \boldsymbol{\Sigma}_{X_j}), \quad (8)$$

where $\boldsymbol{\xi}_t = \left[\boldsymbol{y}_t^\top, 1\right]^\top \in \mathcal{R}^{3D_Y+1}$ is the expanded articulatory feature vector and $\boldsymbol{A}_k \in \mathcal{R}^{3D_X \times (3D_Y+1)}$ is the transform matrix for the k-th mixture of $\lambda^{(G)}$. Fig.1 compares the feature production models used in our baseline and proposed methods. We see that an extra Gaussian mixture sequence m_t is introduced to determine the cross-stream transform matrix for each frame. We can interpret $\zeta_k(t)$ as a weight that varies according to $\boldsymbol{y}_t$, and which changes how each transform matrix is weighted, or "blended" together, according to Eq.(6).

3.2. Model training

To train the HMM parameter set $\{\boldsymbol{A}_k, \boldsymbol{\mu}_{X_j}, \boldsymbol{\Sigma}_{X_j}, \boldsymbol{\mu}_{Y_j}, \boldsymbol{\Sigma}_{Y_j}\}$[1], we substitute Eq.(2), (3), (6), (8) into Eq.(1) and get

$$P(\boldsymbol{X}, \boldsymbol{Y}|\lambda) = \sum_{\boldsymbol{q}} \sum_{\boldsymbol{m}} P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q}, \boldsymbol{m}|\lambda), \quad (9)$$

where

$$P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q}, \boldsymbol{m}|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} \zeta_{m_t}(t) \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{Y_{q_t}}, \boldsymbol{\Sigma}_{Y_{q_t}}) \cdot \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_{m_t} \boldsymbol{\xi}_t + \boldsymbol{\mu}_{X_{q_t}}, \boldsymbol{\Sigma}_{X_{q_t}}). \quad (10)$$

[1] In this work, the covariance matrices $\boldsymbol{\Sigma}_{X_j}$ and $\boldsymbol{\Sigma}_{Y_j}$ of each HMM state are set to be diagonal for simplification.

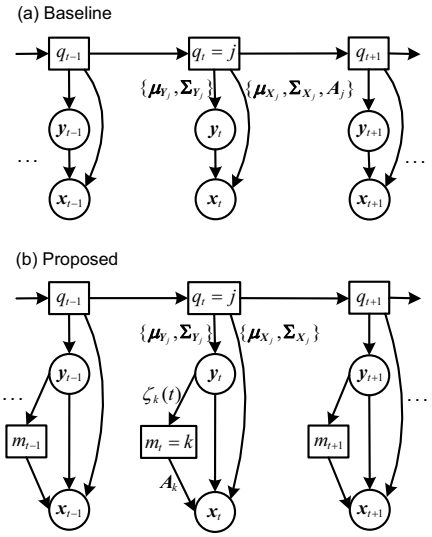

Figure 1: Feature production model for combined acoustic and articulatory modelling in our (a) baseline and (b) proposed methods. The definition of the parameters on the arcs that represent the dependency relationship can be found in Eq.(3), (4) and Eq.(7), (8).

The EM algorithm is adopted to estimate the parameter set that maximizes Eq.(9). The auxiliary function is defined as

$$Q(\lambda, \lambda') = \sum_{\boldsymbol{q}} \sum_{\boldsymbol{m}} P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q}, \boldsymbol{m}|\lambda) \log P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q}, \boldsymbol{m}|\lambda') \quad (11)$$

$$= \sum_{j=1}^{N} \sum_{k=1}^{M} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \Big[\log \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}'_{Y_j}, \boldsymbol{\Sigma}'_{Y_j}) + \log \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}'_k \boldsymbol{\xi}_t + \boldsymbol{\mu}'_{X_j}, \boldsymbol{\Sigma}'_{X_j})\Big] + K, \quad (12)$$

where K is a constant term that is independent of the model parameter set; $\gamma_j(t)$ is the occupancy probability of state j at time t; N is the total number of HMM states.

In order to re-estimate the transform matrix $\boldsymbol{A}'_k$ for each GMM mixture, we set $\partial Q(\lambda, \lambda')/\partial \boldsymbol{A}'_k = 0$ and get

$$\sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \boldsymbol{\Sigma}_{X_j}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_{X_j}) \boldsymbol{\xi}_t^\top = \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \boldsymbol{\Sigma}_{X_j}^{-1} \boldsymbol{A}'_k \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top. \quad (13)$$

This equation can be simplified as

$$\boldsymbol{Z} = \sum_{t=1}^{T} \boldsymbol{V}^{(t)} \boldsymbol{A}'_k \boldsymbol{D}^{(t)}, \quad (14)$$

where

$$\boldsymbol{Z} = \{z_{il}\} = \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \boldsymbol{\Sigma}_{X_j}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_{X_j}) \boldsymbol{\xi}_t^\top, \quad (15)$$

$$\boldsymbol{V}^{(t)} = diag\left\{v_{ii}^{(t)}\right\} = \sum_{j=1}^{N} \gamma_j(t) \boldsymbol{\Sigma}_{X_j}^{-1}, \quad (16)$$

$$\boldsymbol{D}^{(t)} = \left\{ d_{il}^{(t)} \right\} = \zeta_k(t)\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top. \quad (17)$$

According to Eq.(14), each element in $\boldsymbol{Z}$ can be calculated as

$$z_{il} = \sum_{t=1}^{T}\sum_{p=1}^{3D_Y+1} v_{ii}^{(t)} a'_{ip} d_{pl}^{(t)} = \sum_{p=1}^{3D_Y+1} a'_{ip} \sum_{t=1}^{T} v_{ii}^{(t)} d_{pl}^{(t)}. \quad (18)$$

Therefore, the transform matrix $\boldsymbol{A}'_k$ can be updated line by line. For the i-th line,

$$\boldsymbol{a}'_i = \boldsymbol{G}^{(i)-1}\boldsymbol{z}_i, \quad (19)$$

where $\boldsymbol{z}_i = \left[z_{i1}, z_{i2}, ..., z_{i(3D_Y+1)}\right]^\top$; $\boldsymbol{A}'_k = \{a'_{il}\}$ and $\boldsymbol{a}'_i = \left[a'_{i1}, a'_{i2}, ..., a'_{i(3D_Y+1)}\right]^\top$; $\boldsymbol{G}^{(i)} = \{g_{pr}^{(i)}\}$ and $g_{pr}^{(i)} = \sum_{t=1}^{T} v_{ii}^{(t)} d_{pr}^{(t)}$.

The re-estimation formulae for other model parameters can be derived by setting $\partial Q(\lambda, \lambda')/\partial\lambda' = 0$ as

$$\boldsymbol{\mu}'_{X_j} = \frac{\sum_{k=1}^{M}\sum_{t=1}^{T}\gamma_j(t)\zeta_k(t)(\boldsymbol{x}_t - \boldsymbol{A}'_k\boldsymbol{\xi}_t)}{\sum_{t=1}^{T}\gamma_j(t)}, \quad (20)$$

$$\boldsymbol{\Sigma}'_{X_j} = \frac{1}{\sum_{t=1}^{T}\gamma_j(t)}\sum_{k=1}^{M}\sum_{t=1}^{T}\gamma_j(t)\zeta_k(t) \cdot (\boldsymbol{x}_t - \boldsymbol{\mu}'_{X_j} - \boldsymbol{A}'_k\boldsymbol{\xi}_t)(\boldsymbol{x}_t - \boldsymbol{\mu}'_{X_j} - \boldsymbol{A}'_k\boldsymbol{\xi}_t)^\top, \quad (21)$$

$$\boldsymbol{\mu}'_{Y_j} = \frac{\sum_{t=1}^{T}\gamma_j(t)\boldsymbol{y}_t}{\sum_{t=1}^{T}\gamma_j(t)}, \quad (22)$$

$$\boldsymbol{\Sigma}'_{Y_j} = \frac{\sum_{t=1}^{T}\gamma_j(t)(\boldsymbol{y}_t - \boldsymbol{\mu}'_{Y_j})(\boldsymbol{y}_t - \boldsymbol{\mu}'_{Y_j})^\top}{\sum_{t=1}^{T}\gamma_j(t)}. \quad (23)$$

3.3. Parameter generation with articulatory control

Similar to our previous work [6], the maximum likelihood criterion is adopted and only the optimal HMM state sequence is considered in the parameter generation. The generated articulatory features can be modified to control the characteristics of synthetic speech. The detailed steps are introduced as follows:

1) Generate the optimal state sequence $\boldsymbol{q}^*$ using the trained duration distributions [2].
2) Generate the optimal articulatory features $\boldsymbol{Y}^*$. In order to simplify the calculation, only the articulatory stream in the HMM is used, i.e., to maximize

$$P(\boldsymbol{Y}|\lambda, \boldsymbol{q}^*) \approx \prod_{t=1}^{T}\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{Y_{q_t^*}}, \boldsymbol{\Sigma}_{Y_{q_t^*}}). \quad (24)$$

This can be solved using the conventional maximum likelihood parameter generation (MLPG) algorithm [1].
3) Modify the articulatory features by designing function $f(\cdot)$ based on phonetic rules and get $\widehat{\boldsymbol{Y}} = f(\boldsymbol{Y}^*)$.
4) Generate the optimal acoustic features $\boldsymbol{X}^*$ according to the modified articulatory features by maximizing

$$P(\boldsymbol{X}|\widehat{\boldsymbol{Y}}, \lambda, \boldsymbol{q}^*) = \prod_{t=1}^{T}\sum_{k=1}^{M}\zeta_k(t)\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_k\widehat{\boldsymbol{\xi}}_t + \boldsymbol{\mu}_{X_{q_t^*}}, \boldsymbol{\Sigma}_{X_{q_t^*}}). \quad (25)$$

where $\zeta_k(t)$ is calculated based on $\widehat{\boldsymbol{Y}}$. This is an MLPG problem with mixtures of Gaussians at each frame. We can solve it by either considering only the optimal mixture sequence or using an EM-based iterative estimation method [1].

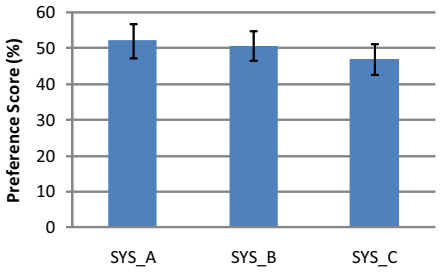

Figure 2: Average preference scores on the naturalness of three system with 95% confidence interval.

4. Experiments

4.1. Database and System Construction

A multi-channel articulatory database was used in our experiments. The acoustic waveform was recorded concurrently with EMA data using a Carstens AG500 electromagnetic articulograph. 1,263 phonetically balanced sentences were read by a male British English speaker. 1,200 sentences were selected for model training. The waveforms were in 16kHz PCM format with 16 bit precision. Six EMA sensors were located at the *tongue dorsum* (T3), *tongue body* (T2), *tongue tip* (T1), *lower lip* (LL), *upper lip* (UL), and *lower incisor* (LI) of the speaker. Each sensor recorded spatial location in 3 dimensions at a 200Hz sample rate: coordinates on the x- (front to back), y- (bottom to top) and z- (left to right) axes (relative to viewing the speaker's face from the front). Because of the very small movements in the z-axis, only the x- and y-coordinates of the six sensors were used in our experiments, making a total of 12 static articulatory features at each frame.

40-order frequency-warped LSPs and an extra gain dimension were derived from the spectral envelop given by STRAIGHT [7] analysis, with a frame shift of 5ms. A 5-state, left-to-right HMM structure with no skips was adopted for the unified acoustic-articulatory modelling. Besides the feature-space transform tying method proposed in this paper, we also evaluated an improved initialization method for the transform matrices in the experiments. Three systems were constructed and compared, which are

SYS_A Baseline method using 100 context-dependent transform matrices. $\boldsymbol{A}_j$ was initialized as a zero matrix for the first iteration of EM re-estimation [6].

SYS_B The same as **SYS_A** except for the initialization strategy for $\boldsymbol{A}_j$. The initial $\boldsymbol{A}_j$ was estimated by fixing $\boldsymbol{\mu}_{X_j} = 0$ for each HMM state in the EM re-estimation.

SYS_C Proposed method using 64 mixtures for model $\lambda^{(G)}$. The same initialization strategy as **SYS_B** was adopted.

4.2. Subjective evaluation

First, a preference test was conducted to compare the naturalness of the three systems. 20 sentences not existing in the training set were selected and synthesized by all three systems. 41 native English listeners took part in the test, which was conducted in listening booths. Fig.2 shows the average preference score of all the listeners. We see that there is no significant difference in naturalness among the three systems.

Then, we carried out a vowel quality modification experiment to evaluate the effectiveness of these three systems in controlling the characteristics of synthetic speech. Five monosyl-

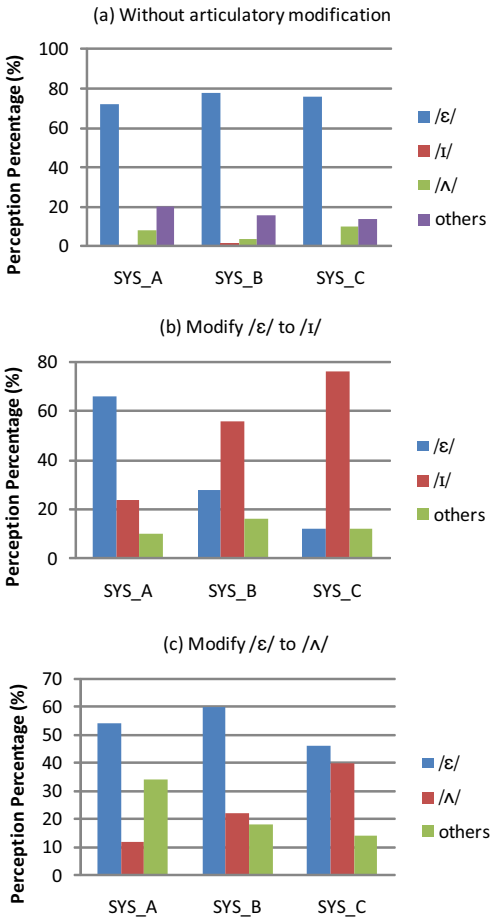

Figure 3: Vowel quality perception results (a) without articulatory modification, (b) modifying /ɛ/ towards /ɪ/, and (c) modifying /ɛ/ towards /ʌ/.

labic words ("*bet*", "*hem*", "*peck*", "*ten*", "*dead*") with vowel /ɛ/ were selected and embedded into the carrier sentence "Now we'll say ... again". We tried to modify the vowel /ɛ/ to be perceived as the vowels /ɪ/ and /ʌ/ by manipulating the generated articulatory trajectory during synthesis. In our previous work [6], we only modified the EMA dimensions corresponding to the height of tongue with a fixed shift (-1.5cm to 1.5cm) to change vowel /ɛ/ to /ɪ/ and /æ/. Although positive results were achieved, the range of modification needed was in fact much larger than the true difference between tongue positions when pronouncing these vowels. In this experiment, the function $f(\cdot)$ in the step 3) of Section 3.3 was designed to replace the generated trajectories of all EMA dimensions for the sentences with vowel /ɛ/ with the ones for vowel /ɪ/ or /ʌ/. We expect such testing scheme can evaluate the speech character controlling ability of different systems when the ideal targets for articulatory modification are known. 10 native English listeners were asked to listen to the three groups (without articulatory modification, modifying /ɛ/ to /ɪ/, and modifying /ɛ/ to /ʌ/) of synthesized samples using each system and to write down the key word in the carrier sentence they heard. Then, we calculated the percentages for how the vowels were perceived as shown in Fig.3.

From Fig.3(a), we see that these three systems have similar dictation correctness for synthetic vowel /ɛ/ without articulatory modification. Figs.3(b) and (c) show the percentage of stimuli perceived as the vowel that is the target of the modification is increased both by the new transform matrix initialization strategy and the proposed feature-space transform tying method. Comparing **SYS_A** with **SYS_C**, the percentage of responses in which the synthetic vowel was correctly perceived as the target vowel increased from 24% to 76% when modifying /ɛ/ to /ɪ/, and from 12% to 40% when modifying /ɛ/ to /ʌ/. Although this improvement is significant, the final performance of vowel modification towards /ʌ/ is not so good. This is because there still exist limitations in our improved model structure when reconstructing acoustic features based on the modified articulatory features. As shown in Fig.1(b), the GMM mixture index m_t can change in response to the articulatory modification. However, this may introduce conflict with the HMM state index q_t and $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ which are still fixed and determined by the context information beforehand.

5. Conclusions

We have presented a feature-space transform tying method for unified acoustic-articulatory modelling, which is used to improve the flexibility of HMM-based parametric speech synthesis. Experimental results have proved the effectiveness of this method in better describing the dependency between acoustic and articulatory feature streams, compared with the conventional method where transform matrices are tied context-dependently. On the other hand, the current model stucture still needs improvement. To further alleviate the restriction of context information and to better model the cross-stream dependency in unified acoustic-articulatory model structure will be the tasks of our future work.

Acknowledgements We thank Phil Hoole of Ludwig-Maximilian University, Munich for his great effort in helping record the EMA data. This work is partially funded by the National Nature Science Foundation of China (Grant No. 60905010). The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and EPSRC grant EP/I027696/1.

6. References

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, vol. 3, 2000, pp. 1315–1318.

[2] K. Tokuda, H. Zen, and A. W. Black, "HMM-based approach to multilingual speech synthesis," in *Text to speech synthesis: New paradigms and advances*, S. Narayanan and A. Alwan, Eds. Prentice Hall, 2004.

[3] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[4] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006: an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[5] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Interspeech 2008*, 2008, pp. 573–576.

[6] ——, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.